Webにおける画像とテキストの融合

AUTHOR(S):

安岡, 孝一

Webにおける画像とテキストの融合

安岡孝一*

1　はじめに

日本の大学における電子図書館構想は、平成4年7月23日の学術審議会答申 [1] 中の「II-6-(2) 大学図書館等の機能強化」に源を発し、平成5年12月16日の学術審議会学術情報資料分科会学術情報部会報告 [2] を経て、平成8年7月29日の学術審議会建議 [3] で一定の方向が示された。この建議から6年、各図書館の目録情報の入力は非常に促進され、文献の所在はインターネットでたちどころにわかるようになってきている。

しかし、この建議の中心論題である所蔵資料そのものの電子化については、現在もお寒い状況が続いている。各図書館は「貴重書画像データベース」などと題して、所蔵資料の画像をWWWページに漫然と並べているだけである。電子化の最終目的である全文検索はおろか、内容に関する電子的な目次も索引も準備されていない。つまるところ、これらの資料を「電子的に読む」という視点が決定的に欠落しており、所蔵資料の美術館に堕しているのである。図書館は、書物を「読む」場であって「見る」場ではない。それは電子図書館でも同じである。

本稿では、この問題に対し、「文字列検索が可能な画像フォーマット」というコンセプトの元に、筆者のこれまでの研究成果も踏まえて、3種類の画像フォーマットを紹介する。また、これらの画像フォーマットの得失を、ファイルの大きさ、検索のしやすさ、複数ページへの対応、などの観点から論じる。これらの画像フォーマットに対する研究と実践を進めることにより、現在の美術館化した電子図書館に少しでも警鐘を鳴らすことができれば幸いである。

2　透明テキスト付き画像フォーマット

この章では、文字列検索が可能な画像フォーマットとして、透明テキスト付きPDF、透明テキスト付きSVG、テキストビハインドDjVuという、3種類の透明テキスト付き画像フォーマットについて述べる。

2.1　透明テキスト付きPDF

筆者がこれまでに研究してきた「文字列検索が可能な画像PDF」の成果 [4, 6] の一部は、Adobe社のKen Lundeとのコラボレーションによって「透明テキスト付きPDF」という形でPDF-1.4 [8] に採用され、日本国内の多くのOCRソフトウェアが、この機能を取り入れるようになってきている。「透明テキスト付きPDF」の実現方法は、現実には2種類ある。画像の上に透明なテキストを重ねる方法と、白い文字で書かれたテキストの上に画像を重ねる方法である。いずれも、最新の

*京都大学人文科学研究所附属漢字情報研究センター

Adobe Acrobat Reader でテキスト検索をおこなうと、当該テキストと重なっている部分の画像が反転する。しかし、旧版の Adobe Acrobat Reader の場合は、後者の方しかテキスト検索がおこなえない。この点から、後者の「白い文字で書かれたテキストの上に画像を重ねる方法」がよく用いられており、本稿でもこちらを紹介することにする。

2.2 透明テキスト付き SVG

SVG (Scalable Vector Graphics) において、文字と画像を重ねて配置することで画像中の文字列をテキスト検索できる方法であり、守岡知彦の研究成果の一つである [9]。SVG のフォーマット [7] そのものは、W3C によって標準化されているオープンな規格であり、多くのビューワが実現されている。ただし、漢字検索に対応しているものは、現時点では Adobe SVG Viewer 3.0 くらいのものである。なお守岡は、テキストの上に画像を重ねる方法を取っているが、本稿では、画像の上に透明なテキストを重ねる方法を紹介することにする。

2.3 テキストビハインド DjVu

AT&T の Léon Bottou らによって開発されている DjVu は、DjVu 3.0 [5] において「テキストビハインド (Hidden Text)」機能が追加され、3.0 以前の単なる圧縮率が高いだけの画像フォーマットから、テキスト検索可能な画像フォーマットへの脱皮が図られた。DjVu そのものはオープンソースとなってはいるが、現時点では LizardTech Software 社がほぼ独占的に販売をおこなっており、ビューワも LizardTech Software 社 (国内ではイメージリアリティ社) からダウンロードできるものがほぼ唯一である。本稿では、ビューワとしてはイメージリアリティ社のものを用いるが、DjVu 制作ソフトウェアとしては、オープンソースの djvulibre 3.5.10 を Solaris 上で用いることにする。

3 透明テキスト付き画像フォーマットの比較

袁安碑の jpg 画像 (図 1、1159×2374 ピクセル、666260 バイト) である `enanhi.jpg` を元に、釈文を埋め込んだ透明テキスト付き PDF、透明テキスト付き SVG、テキストビハインド DjVu の作成をおこなった。

透明テキスト付き PDF ファイル `enanhi.pdf` は、付録 A に示す PDF を手作業で (つまりはテキストエディタで) 作成した。Adobe Photoshop で袁安碑の jpg に対して透明なテキストを重ね、それを PDF で出力する方が作業的には楽だが、Adobe Photoshop の出力する PDF はファイルサイズが巨大になってしまいやすいことから、あえて手作業を選んだ。なお、漢字部分の文字コードは、日本語 EUC である。

透明テキスト付き SVG ファイル `enanhi.svg` は、`svg10.dtd` にしたがい、付録 B に示す SVG をやはり手作業で作成した。文字コードは UTF-8 である。

テキストビハインド DjVu は、まず djvulibre の `c44` によって `enanhi.jpg` から `enanhi.djvu` を作成し、さらに手作業で書いた付録 C の DjVuXML ファイル

enanhi.xml を、djvulibre の djvuxmlparser を使って enanhi.djvu に取り込んだ。なお、enanhi.xml の文字コードも UTF-8 である。

3.1 ファイルサイズの比較

ファイルサイズは、それぞれ以下のようになった。

透明テキスト付き PDF	668340 バイト
透明テキスト付き SVG	1827+666260=668087 バイト
テキストビハインド DjVu	566250 バイト

透明テキスト付き SVG は、enanhi.svg そのもののサイズは 1827 バイトだが、これとは別に enanhi.jpg が必要なことから、合計バイト数としている。テキストビハインド DjVu に関しては、c44 で作成した時点の enanhi.djvu は 565679 バイト、enanhi.xml のサイズは 2006 バイトだったが、djvuxmlparser の結果 enanhi.djvu が 566250 バイトとなったことから、これを示している。

ファイルサイズを比較すると、透明テキスト付き PDF と透明テキスト付き SVG は大差ないが、これらに比べテキストビハインド DjVu は約 8%小さくなっている。圧縮率の高さが効いているということであろう。

3.2 検索動作の比較

検索動作の比較は、いきおいビューワの作りの良し悪しの比較になりかねないのだが、ここではフォーマットに依存する検索動作の違いを比較したい。なお検索はいずれも、Microsoft Windows Me 上の Internet Explorer 5.5 でのプラグインを使用することにする。

enanhi.pdf の Adobe Acrobat Reader 5.1.0 による表示を図 2 に、enanhi.svg の Adobe SVG Viewer 3.0 Build 76 による表示を図 3 に、enanhi.dvju の DjVu Browser Plug-in 3.6.2 による表示を図 4 に、それぞれ示す。いずれも「十三年」を検索した際の表示である。

一見してわかるのは、テキストビハインド DjVu の検索結果の表示範囲が、実際の文字列より広くなっている、という点である。これは、DjVu の検索が、DjVuXML での WORD タグに対しておこなわれるためである。すなわち「十三年」という検索が、付録 C の DjVuXML で言えば

```
<WORD coords="676,333,597,1769">十三年十二月丙辰</WORD>
```

の部分にマッチングしていることから、(676,333)-(597,1769) の範囲、つまり「十三年十二月丙辰」の部分が反転表示されてしまうのである。では、これを防ぐためにたとえば

```
<WORD coords="676,333,597,467">十</WORD>
<WORD coords="676,467,597,602">三</WORD>
<WORD coords="676,602,597,744">年</WORD>
```

```
<WORD coords="676,744,597,879">十</WORD>
<WORD coords="676,1176,597,1325">二</WORD>
<WORD coords="676,1325,597,1468">月</WORD>
<WORD coords="676,1468,597,1626">丙</WORD>
<WORD coords="676,1626,597,1769">辰</WORD>
```

と定義すればどうなるだろう。実は、これでは「十三年」にはマッチングしなくなってしまう。DjVu のテキストマッチングは、WORD 間に空白があることを想定しているため、上記の一文字ずつの WORD タグだと、「十 三 年」でしかマッチングがおこらなくなってしまうのである。この意味で、DjVu のテキストマッチングと漢字列検索との間は、微妙に齟齬をきたしていると言えよう。

4 電子図書館への対応

透明テキスト付き画像フォーマットを、実際に電子図書館で用いることを考えてみよう。一般の資料は拓本とは違い、複数のページを有するのが通常である。しかも、複数ページの資料だからといって、必ずしも最初のページから順に読むとは限らず、文書内部での検索や、目次や索引による他のページへのジャンプ、といった機構が準備されていなければならない。この章では、複数ページの資料を「電子的に読む」という視点から、透明テキスト付き画像フォーマットを比較していくことにしよう。

4.1 複数ページに渡る検索

1 つの PDF に全ページをまとめてさえおけば、Adobe Acrobat Reader は文書全体に渡る文字列検索が可能である。ただし、ページ数が増えれば、ファイルサイズは非常に巨大なものとなるのは覚悟しなければならない。

SVG には、そもそも複数ページという概念がない。各ページをバラバラに SVG で記述し、その間にリンクを張るしかなく、複数ページに渡る検索は Viewer 側では実現不可能である。

DjVu では複数ページの資料に対して、全ページを 1 つのファイルにまとめる Bundled 形式と、各ページをバラバラのファイルで扱う Indirect 形式とを、ファイルフォーマットとして準備している。しかも、いずれのフォーマットに対しても、DjVu Browser Plug-in は、ページめくりや複数ページに渡る検索が可能となっており、しかも Indirect 形式においては、各ページは必要に応じて Browser に読み込まれるようになっている。

4.2 目次と索引

PDF では、目次や索引から当該ページにジャンプする機構は、PDF ファイル間で可能である。したがって目次や索引は、本文と同じ PDF に含まれていてもかまわないし、本文と別の PDF でもかまわない。

SVG においては、HTML や SVG のリンクによって、目次や索引を一応記述できる。もちろんリンク先は、バラバラに記述された各ページとなる。

DjVu には、目次や索引を記述するためのリンク機構がない。ありていに言えば、DjVu 内部から他の URL へのジャンプはおこなえない。したがって、目次や索引は外部に HTML 等で記述しておき、そのリンク先は Indirect 形式 DjVu の当該ページ、というやり方になる。

5 おわりに

美術館化した電子図書館に警鐘を鳴らすべく、3つの透明テキスト付き画像フォーマットを紹介するとともに、それらの例を実際に手作業で制作し、その得失について論じた。

英語などの「単語を分かち書きする言語」に対しては、Indirect 形式 DjVu が圧倒的に良いが、漢字に関しては検索処理という点で、透明テキスト付き PDF に軍配があがる。ただ、ページ数が増大した場合、PDF はページ数に比例してファイルが大きくなってしまうことから、Indirect 形式 DjVu の軽さは捨てがたい。この点から、DjVu の検索機能を日本語や中国語にフィットさせる研究が、今後は必要となるように思われる。あるいは、複数ファイルに分割された PDF をまとめて扱うような Reader や、DjVu 内にリンクを埋め込む機能などの開発も、今後重要となっていくだろう。

参考文献

[1] 21 世紀を展望した学術研究の総合的推進方策について (答申), 学術審議会 (1992 年 7 月).

[2] 大学図書館機能の強化・高度化の推進について (報告), 学術審議会学術情報資料分科会学術情報部会 (1993 年 12 月).

[3] 大学図書館における電子図書館的機能の充実・強化について (建議), 学術審議会 (1996 年 7 月).

[4] Koichi Yasuoka and Tokio Takata: Digital Rubbings — Their Past and Future, 2001 Pacific Neighborhood Consortium Proceedings (January 2001), ECAI Rubbings Work Session.

[5] Yan Le Cun, Léon Bottou, Andrei Erofeev, Patrick Haffner, Bill Riemers: DjVu Document Browsing with On-Demand Loading and Rendering of Image Components, IS&T/SPIE's 13th Annual Symposium on Electronic Imaging: Science and Technology — Internet Imaging II (January 2001), Paper No.4311-02.

[6] 安岡孝一: フォント埋め込みによる外字手法, 京都大学大型計算機センター第 67 回研究セミナー報告 (2001 年 3 月), pp.3-12.

[7] Scalable Vector Graphics (SVG) 1.0 Specification, W3C Recommendation `http://www.w3.org/TR/SVG/` (September 2001).

[8] Adobe Systems Incorporated: PDF Reference third edition — Adobe Portable Document Format Version 1.4, Addison-Wesley (December 2001).

[9] 守岡知彦: ポスト文字コード時代の文書処理技術に関する展望, 全国文献・情報センター人文社会科学学術セミナーシリーズ, No.12 (2002 年 11 月), pp.59-70.

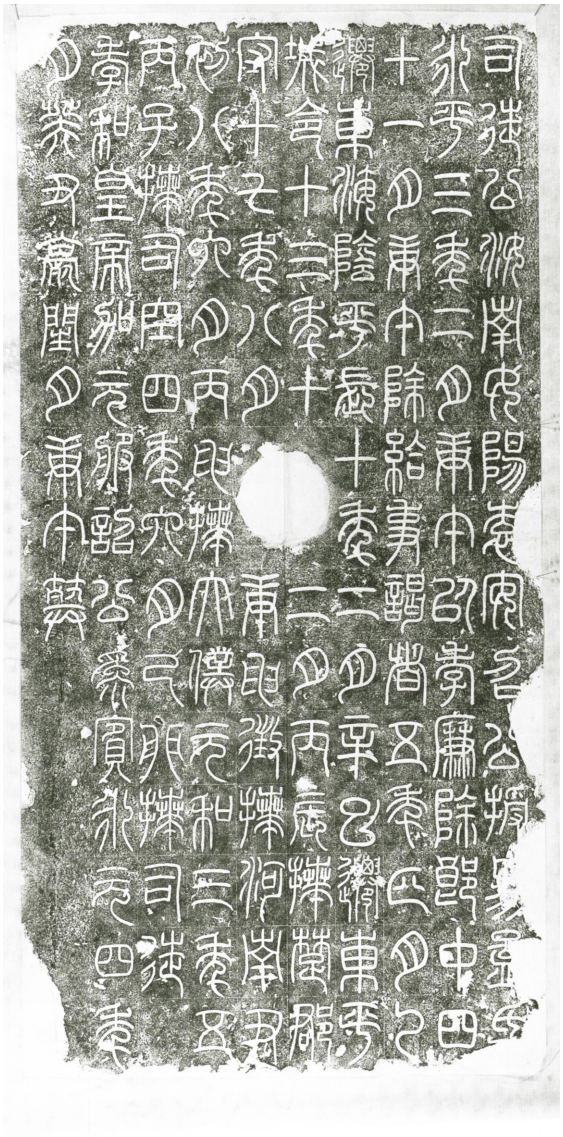

図 1: 袁安碑の jpg 画像 `enanhi.jpg`

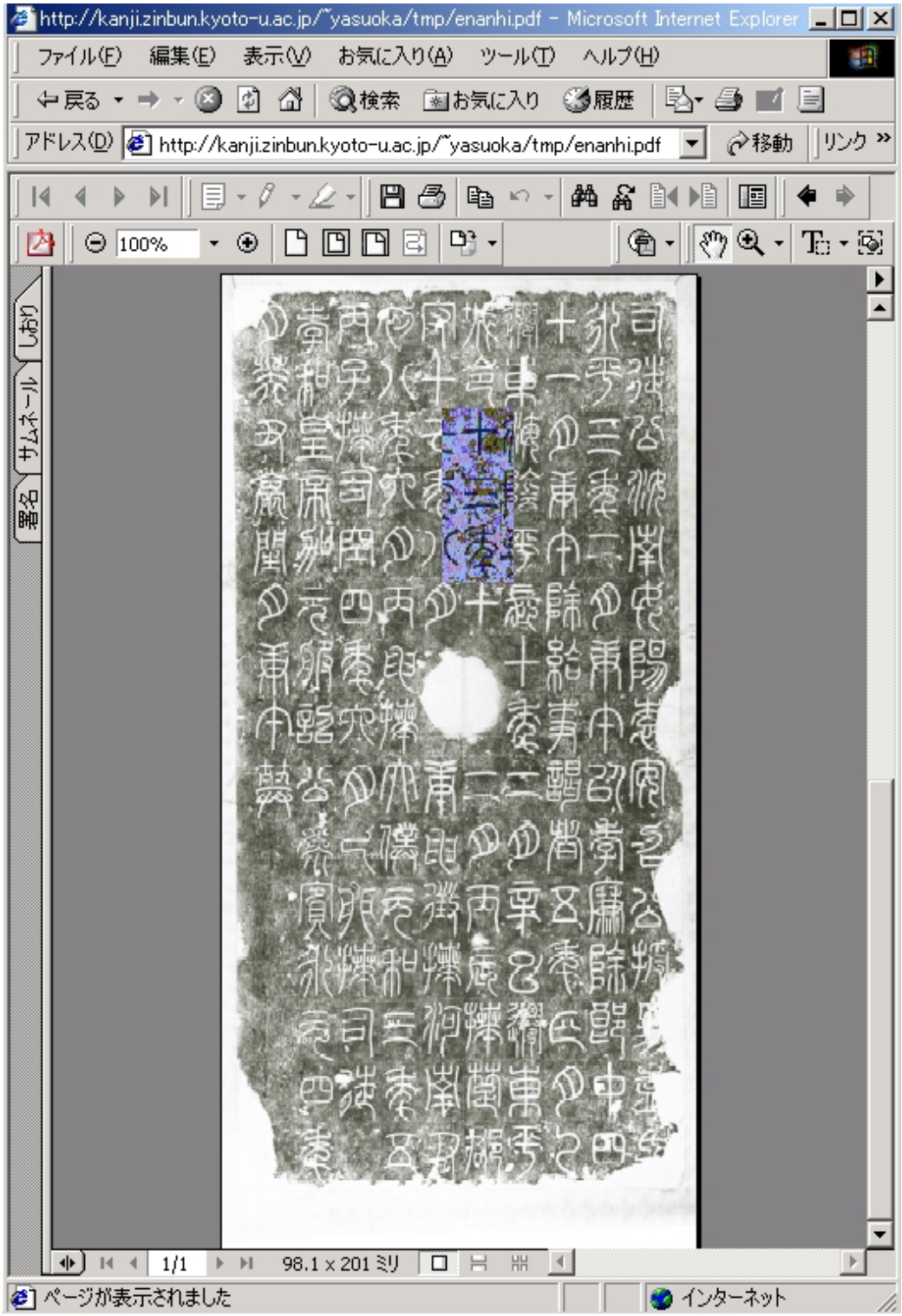

図 2: 透明テキスト付き PDF で「十三年」を検索

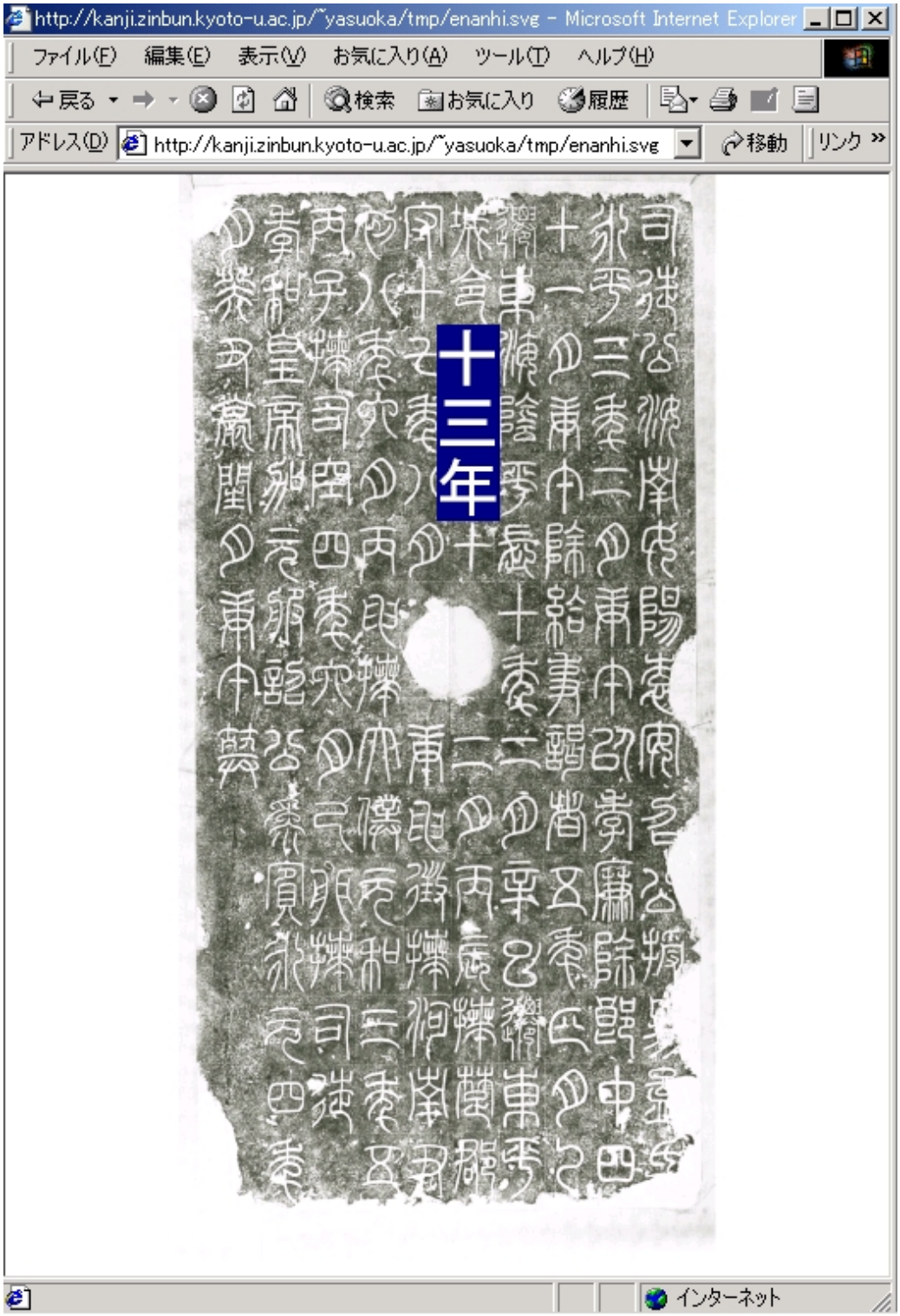

図 3: 透明テキスト付き SVG で「十三年」を検索

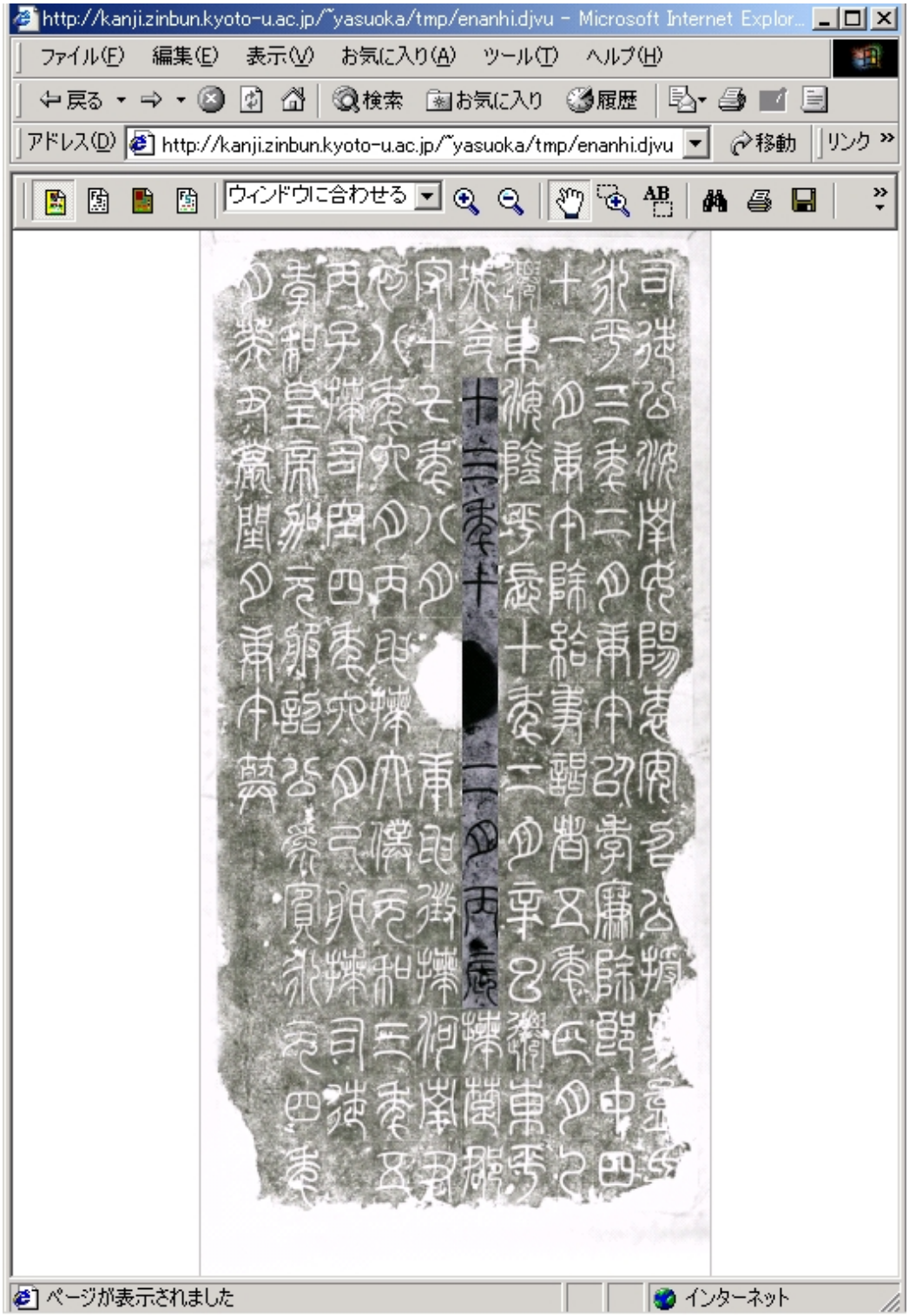

図 4: テキストビハインド DjVu で「十三年」を検索

付録 A 透明テキスト付き PDF(袁安碑)

```
%PDF-1.3
1 0 obj << /Type /Catalog /Pages 2 0 R >> endobj
2 0 obj << /Type /Pages /Kids [3 0 R] /Count 1 >> endobj
3 0 obj << /Type /Page /MediaBox [0 0 278.16000 569.76000] /Parent 2 0 R
/Resources << /ProcSet [/PDF /Text /ImageC] /Font << /F1 6 0 R >>
/XObject << /R1 5 0 R >> >> /Contents 4 0 R >> endobj
4 0 obj << /Length 932 >> stream
q 0.24 0 0 0.24 0 0 cm 1 g BT
/F1 140 Tf 1040 2328 Td (司徒公汝南女陽袁安) Tj
/F1 147 Tf 10 -1279 Td (召公授易孟氏学) Tj
/F1 140 Tf -115 1279 Td (永平三年二月庚午以) Tj
/F1 147 Tf 10 -1279 Td (孝廉除郎中四年) Tj
/F1 140 Tf -115 1279 Td (十一月庚午除給事謁) Tj
/F1 147 Tf 10 -1279 Td (者五年四月乙□) Tj
/F1 140 Tf -105 1279 Td (遷東海陰平長十年二) Tj
/F1 147 Tf 10 -1279 Td (月辛巳遷東平任) Tj
/F1 140 Tf -120 1279 Td (城令十三年十) Tj
/F1 147 Tf 20 -1128 Td (二月丙辰拝楚郡太) Tj
/F1 140 Tf -120 1128 Td (守十七年八月) Tj
/F1 147 Tf 30 -1128 Td (庚申徴拝河南尹建) Tj
/F1 140 Tf -125 1128 Td (初八年六月丙申拝太) Tj
/F1 147 Tf 10 -1279 Td (僕元和三年五月) Tj
/F1 140 Tf -115 1279 Td (丙子拝司空四年六月) Tj
/F1 147 Tf 10 -1279 Td (己卯拝司徒) Tj
/F1 140 Tf -110 1279 Td (孝和皇帝加元服詔公) Tj
/F1 147 Tf 10 -1279 Td (為賓永元四年三) Tj
/F1 140 Tf -110 1279 Td (月癸丑斃閏月庚午葬) Tj
ET Q
q 278.16000 0 0 569.76000 0 0 cm /R1 Do Q
endstream
endobj
6 0 obj << /Type /Font /Subtype /Type0 /BaseFont /Ryumin-Light-EUC-V
/Encoding /EUC-V /DescendantFonts [7 0 R] >> endobj
7 0 obj << /Type /Font /Subtype /CIDFontType0 /BaseFont /Ryumin-Light
/FontDescriptor 8 0 R /CIDSystemInfo << /Registry (Adobe)
/Ordering (Japan1) /Supplement 0 >> /DW 1000 >> endobj
8 0 obj << /Type /FontDescriptor /Ascent 723 /CapHeight 709 /Descent -241
/Flags 6 /FontBBox [-170 -331 1024 903] /FontName /Ryumin-Light
/ItalicAngle 0 /StemV 69 >> endobj
5 0 obj << /Subtype /Image /ColorSpace /DeviceRGB /Width 1159
/Height 2374 /BitsPerComponent 8 /Filter /DCTDecode /Length 666260
>> stream
```

[袁安碑 jpg データ]

```
endstream
endobj
xref 0 9
0000000000 65535 f
0000000009 00000 n
0000000058 00000 n
```

```
0000000115 00000 n
0000000308 00000 n
0000001767 00000 n
0000001290 00000 n
0000001411 00000 n
0000001594 00000 n
trailer << /Size 9 /Root 1 0 R >>
startxref
668184
%%EOF
```

付録B 透明テキスト付きSVG(袁安碑)

```
<?xml version="1.0" encoding="utf-8" standalone="no"?>
<!DOCTYPE svg PUBLIC "-//W3C//DTD SVG 20010904//EN"
"http://www.w3.org/TR/2001/REC-SVG-20010904/DTD/svg10.dtd">
<svg width="100%" height="100%" viewBox="0 0 1159 2374">
<image x="0" y="0" width="1159" height="2374" xlink:href="enanhi.jpg"/>
<text style="writing-mode:tb; fill-opacity:0">
<tspan x="1040" y="46" style="font-size:140"
>司徒公汝南女陽袁安</tspan><tspan x="1050" y="1325" style="font-size:147"
>召公授易孟氏学</tspan><tspan x="935" y="46" style="font-size:140"
>永平三年二月庚午以</tspan><tspan x="945" y="1325" style="font-size:147"
>孝廉除郎中四年</tspan><tspan x="830" y="46" style="font-size:140"
>十一月庚午除給事謁</tspan><tspan x="840" y="1325" style="font-size:147"
>者五年四月乙□</tspan><tspan x="735" y="46" style="font-size:140"
>遷東海陰平長十年二</tspan><tspan x="745" y="1325" style="font-size:147"
>月辛巳遷東平任</tspan><tspan x="625" y="46" style="font-size:140"
>城令十三年十</tspan><tspan x="645" y="1174" style="font-size:147"
>二月丙辰拝楚郡太</tspan><tspan x="525" y="46" style="font-size:140"
>守十七年八月</tspan><tspan x="555" y="1174" style="font-size:147"
>庚申徴拝河南尹建</tspan><tspan x="430" y="46" style="font-size:140"
>初八年六月丙申拝太</tspan><tspan x="440" y="1325" style="font-size:147"
>僕元和三年五月</tspan><tspan x="325" y="46" style="font-size:140"
>丙子拝司空四年六月</tspan><tspan x="335" y="1325" style="font-size:147"
>己卯拝司徒</tspan><tspan x="225" y="46" style="font-size:140"
>孝和皇帝加元服詔公</tspan><tspan x="235" y="1325" style="font-size:147"
>為賓永元四年三</tspan><tspan x="125" y="46" style="font-size:140"
>月癸丑斃閏月庚午葬</tspan>
</text>
</svg>
```

付録C テキストビハインド DjVuの定義XML(袁安碑)

```
<?xml version="1.0" encoding="utf-8" standalone="no"?>
<!DOCTYPE DjVuXML PUBLIC "-//W3C//DTD DjVuXML 1.1//EN"
"file:/usr/local/djvu/share/djvu/pubtext/DjVuXML-s.dtd">
<DjVuXML>
<HEAD>enanhi.djvu</HEAD>
```

<BODY>
<OBJECT data="enanhi.djvu" type="image/x.djvu" width="1159" height="2374">
<PARAM name="DPI" value="300"/>
<PARAM name="GAMMA" value="2.200000"/>
<HIDDENTEXT>
<WORD coords="1000,44,1103,1618">司徒公汝南女陽袁安召公</WORD>
<WORD coords="1103,1618,1016,2187">授易孟氏学</WORD>
<WORD coords="893,53,984,1178">永平三年二月庚午</WORD>
<WORD coords="984,1178,901,2076">以孝廉除郎中</WORD>
<WORD coords="901,2076,996,2218">四年</WORD>
<WORD coords="870,44,791,737">十一月庚午</WORD>
<WORD coords="791,737,890,1475">除給事謁者</WORD>
<WORD coords="890,1475,799,2218">五年四月乙</WORD>
<WORD coords="775,48,692,880">遷東海陰平長</WORD>
<WORD coords="692,880,791,1764">十年二月辛巳</WORD>
<WORD coords="791,1764,700,2209">遷東平任</WORD>
<WORD coords="680,44,589,333">城令</WORD>
<WORD coords="676,333,597,1769">十三年十二月丙辰</WORD>
<WORD coords="597,1769,692,2227">拝楚郡</WORD>
<WORD coords="486,44,585,182">太守</WORD>
<WORD coords="577,195,486,1480">十七年八月庚申</WORD>
<WORD coords="486,1480,589,2222">徴拝河南尹</WORD>
<WORD coords="470,35,391,1035">建初八年六月丙申</WORD>
<WORD coords="391,1035,478,1475">拝太僕</WORD>
<WORD coords="478,1475,399,2209">元和三年五月</WORD>
<WORD coords="359,44,276,320">丙子</WORD>
<WORD coords="276,320,375,742">拝司空</WORD>
<WORD coords="375,742,292,1622">四年六月己卯</WORD>
<WORD coords="292,1622,387,2076">拝司徒</WORD>
<WORD coords="178,35,272,1031">孝和皇帝加元服</WORD>
<WORD coords="272,1031,185,1631">詔公為賓</WORD>
<WORD coords="185,1631,276,2205">永元四年</WORD>
<WORD coords="166,44,75,609">三月癸丑斃</WORD>
<WORD coords="75,609,178,1347">閏月庚午葬</WORD>
</HIDDENTEXT>
</OBJECT>
</BODY>
</DjVuXML>